Tackling Challenges in Research Data Management

Alex Ball

21 January 2013

Hello, my name is Alex Ball and I work for the Digital Curation Centre (DCC). In case you haven't heard of us, ¶ we are a collaboration between

- University of Edinburgh
- HATII, University of Glasgow
- UKOLN, University of Bath, where I am based.

We identify, compile and disseminate best practice; and provide advice, support and training on research data management at the institutional level through our programme of institutional engagements.

Normally at this point I would have a slide or two about the importance of research data management, but as we've heard a lot about that already today, let's jump straight in, looking at the challenges facing institutions and researchers. ¶

We'll begin at the institutional level, where we're concerned with technical infrastructure, support structures and so on. Then we'll move down to the project level and work through the research lifecycle, from planning, through collection and analysis, to dissemination. And because I don't have time to go into detail, I'll finish with pointers to more information. ¶

1 Planning institutional readiness for research data management

Until very recently, institutions (in the UK at least) did not have to care very much about research data management. The disciplines that really cared about data – typically those making unrepeatable observations or using breathtakingly expensive equipment – had made their own arrangements with dedicated data centres and repositories. And it was these centres that provided advice and support to researchers.

But when the UK signed up to the OECD *Declaration on Access to Data from Public Funding* in 2004 it pushed the issue of open data up the political agenda, and research funders started looking harder at what was happening to the data they had paid for. Meanwhile,

the closure of the Arts and Humanities Data Service in 2008 and the 'climategate' scandal of 2009–10 were warnings for institutions to take data management more seriously. The game changer came in 2011, though, when the EPSRC published their data policy. ¶ The other major funders laid funding conditions on applicants. The EPSRC laid them on institutions:

> 'EPSRC expects all those [research organisations] it funds to have developed a *clear roadmap* to align their policies and processes with EPSRC's expectations by 1st May 2012, and to be *fully compliant* with these expectations by 1st May 2015.'

Here are EPSRC's expectations (paraphrased):

1. *Research organisations (ROs)* to raise awareness of data sharing responsibilities and issues.
2. Publications should link to underlying data.
3. *ROs* must keep track of their research datasets and requests for them.
4. Born-analogue data must also be shareable on request.
5. *ROs* must provide open, online catalogues of their data; digital data must be given a robust ID.
6. Access restrictions should be clear and justified.
7. *ROs* must provide access to data for 10 years from last access.
8. *ROs* must curate their research data.
9. *ROs* must pay for this from their existing public funding streams.

So what does an institution need to do to avoid being blacklisted by the EPSRC? Well, by now they should have a roadmap and the DCC has helped a lot of institutions to write theirs. But getting down to specifics, it can be tough to know where to begin. But there are tools to help. ¶

To get a handle on what data an institution holds, there's the Data Asset Framework, or DAF. This was originally intended as a methodology for assembling an inventory of data assets, based on desk research, interviews and questionnaires. But what its users found really valuable were the insights it gave them into the state of current practice and the scale and variety of the data assets out there in the wild. ¶

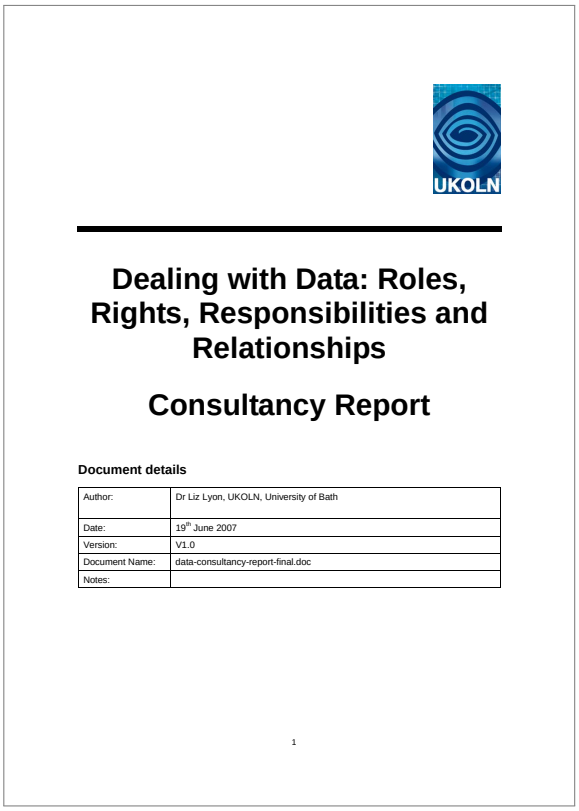
Taking things one step further there's CARDIO, a sort of health check for institutional RDM. It comes in two flavours. The first is very quick and easy: ten multiple-choice questions that guide you through the main areas of data management and invite you to reflect on how well you're doing in each of them. ¶ The second is rather more thorough. It invites you and other stakeholders to assess the institution's performance in 30 different areas, and provides facilities for getting a consensus view and formulating a concrete action plan. This might include writing policies on IPR or risk management, rethinking how IT facilities are financed, providing new infrastructure such as a data catalogue or repository, or providing data management training.

Co-incidentally, these are all things that principal investigators need to know about when they are writing funding proposals, because they come up as part of the data management plan. ¶

2 Planning for research data management at the project level

¶ Data Management Plans have been around in one form or another since at least 1996, but only started showing up as a required part of funding proposals in the mid-2000s. Leading the way were NERC, MRC and the Wellcome Trust. Liz Lyon saw this when writing her *Dealing with Data* report and thought they were a jolly good idea.

> **Recommendation 9.** Each funded research project, should submit a structured Data Management Plan for peer-review as an integral part of the application for funding. — Liz Lyon (2007), *Dealing with Data: Roles, Rights, Responsibilities and Relationships* (University of Bath)



UKOLN

Dealing with Data: Roles, Rights, Responsibilities and Relationships

Consultancy Report

Document details

Author:	Dr Liz Lyon, UKOLN, University of Bath
Date:	19th June 2007
Version:	V1.0
Document Name:	data-consultancy-report-final.doc
Notes:	

1

Why? ¶ Writing and using a Data Management Plan helps

- to co-ordinate the actions of data stakeholders
- to ensure all necessary tasks are accomplished
- to ensure data are properly curated
- with releasing data in a timely fashion
- with sharing data as openly as possible
- with preserving data for future use

Just as research has a lifecycle, so does a data management plan. The first stage is at the point of bidding for funding. Here the purpose of the plan is demonstrate that the applicant has thought about data issues, so they won't waste time collecting data that already exist, and the new data they produce will be usable and shareable. Once the bid has been accepted, the plan needs to be firmed up to reflect the practical realities of the research. Currently only NERC mandates this stage but it needs doing. It's also a good idea to review it periodically throughout the project to make sure that it is being followed, and make any necessary adjustments.

Towards the end of the project, the DMP becomes a useful part of the data management record, which can be handed over to a data centre or repository as evidence for the provenance of the data. The data centres and repositories will themselves have data management plans that are mostly focused on curation and long-term preservation.

The DCC runs a service called DMP Online, which allows people to:

1. create, store and update Data Management Plans

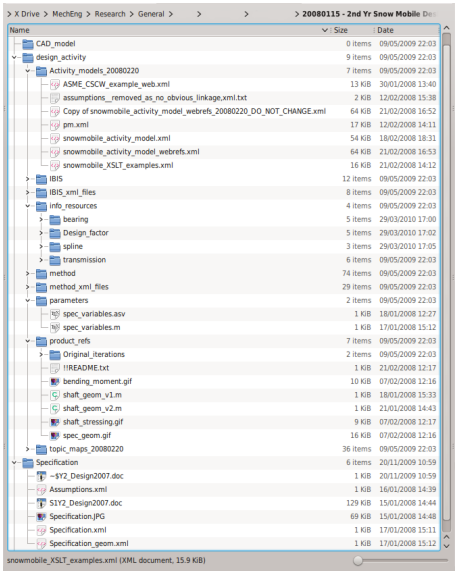


Figure 1: Files and directories relating to the 2nd Year Snowmobile Design Task data case from the KIM Project

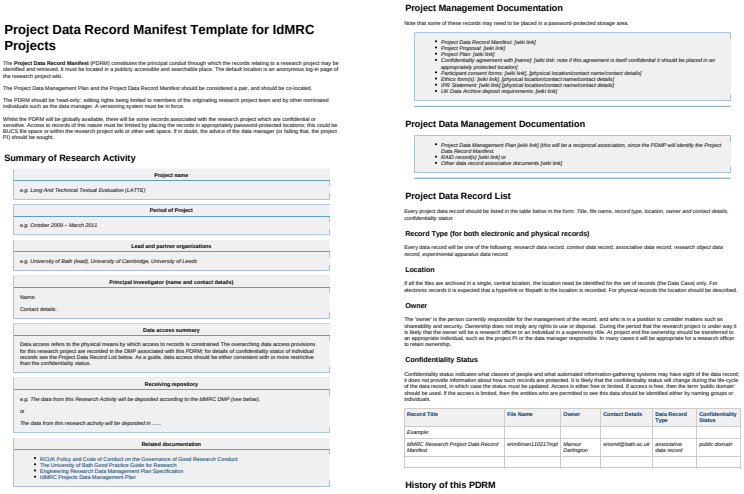
2. meet both institutional and funders' data-related requirements, all in one go
3. receive specific guidance from funders and institutions
4. export Data Management Plans in various formats

We talk to institutions a lot about this tool, and a couple of issues come up. First, institutions find these plans very interesting because they reveal researchers' real data management needs; they're very useful for planning storage provision, for example. Competing with that are concerns over privacy. Some projects are so sensitive that even the data management plan might give away too much information. For the Department of Mechanical Engineering in Bath, we decided we should have a public space for listing projects and their DMPs, but if DMPs were sensitive, they should be kept in a secure space, and the public page would simply point to that location and explain who was allowed to access it. We also encouraged researchers to provide a redacted version of sensitive documents in the public space wherever possible. All this was done in the context of a project called REDm-MED, which brings me onto our next challenge. ¶

3 Monitoring data-related activity

In REDm-MED and its precursor project ERIM, we looked at the data produced by engineering researchers across a range of projects, looking for common features. ¶ What we actually found was incredible diversity: it seemed every project was working with a different set of formats and using a different workflow. It's unlikely that anyone coming to directory of data like this (Figure 1), even the researchers themselves a few months on, would know what it all means and how it fits together.

So, in ERIM and REDm-MED we decided the way to solve this would be to create a Project Record Manifest (Figure 2).



Project Data Record Manifest Template for IdMRC Projects

The **Project Data Record Manifest** (PDRM) constitutes the principal conduit through which the records relating to a research project may be identified and retrieved. It must be located in a publicly accessible and searchable place. The default location is an anonymous log-in page of the research project wiki.

The Project Data Management Plan and the Project Data Record Manifest should be considered a pair, and should be co-located.

The PDRM should be 'read-only', editing rights being limited to members of the originating research project team and by other nominated individuals such as the data manager. A versioning system must be in force.

Whilst the PDRM will be globally available, there will be some records associated with the research project which are confidential or sensitive. Access to records of this nature must be limited by placing the records in appropriately password-protected locations; this could be BUCS file space or within the research project wiki or other web space. If in doubt, the advice of the data manager (or failing that, the project PI) should be sought.

Summary of Research Activity

Project name
e.g. Long And Technical Textual Evaluation (LATTE)
Period of Project
e.g. October 2009 – March 2011
Lead and partner organizations
e.g. University of Bath (lead), University of Cambridge, University of Leeds
Principal Investigator [name and contact details]
Name: Contact details:
Data access summary
Data access refers to the physical means by which access to records is constrained The overarching data access provisions for this research project are recorded in the DMP associated with this PDRM; for details of confidentiality status of individual records see the Project Data Record List below. As a guide, data access should be either consistent with or more restrictive than the confidentiality status.
Receiving repository
e.g. The data from this Research Activity will be deposited according to the IdMRC DMP (see below). or *The data from this research activity will be deposited in ……*
Related documentation
• RCUK Policy and Code of Conduct on the Governance of Good Research Conduct • The University of Bath Good Practice Guide for Research • Engineering Research Data Management Plan Specification • IdMRC Projects Data Management Plan

Project Management Documentation

Note that some of these records may need to be placed in a password-protected storage area.

- *Project Data Record Manifest: [wiki link]*
- *Project Proposal: [wiki link]*
- *Project Plan: [wiki link]*
- *Confidentiality agreement with [name]: [wiki link: note if this agreement is itself confidential it should be placed in an appropriately protected location]*
- *Participant consent forms: [wiki link], [physical location/contact name/contact details]*
- *Ethics form(s): [wiki link], [physical location/contact name/contact details]*
- *IPR Statement: [wiki link] [physical location/contact name/contact details]*
- *UK Data Archive deposit requirements: [wiki link]*

Project Data Management Documentation

- *Project Data Management Plan [wiki link] [this will be a reciprocal association, since the PDMP will identify the Project Data Record Manifest.*
- *RAID record(s): [wiki link] or*
- *Other data record associative documents [wiki link]*

Project Data Record List

Every project data record should be listed in the table below in the form: Title, file name, record type, location, owner and contact details, confidentiality status

Record Type (for both electronic and physical records)

Every data record will be one of the following: research data record, context data record, associative data record, research object data record, experimental apparatus data record

Location

If all the files are archived in a single, central location, the location need be identified for the set of records (the Data Case) only. For electronic records it is expected that a hyperlink or filepath to the location is recorded. For physical records the location should be described.

Owner

The 'owner' is the person currently responsible for the management of the record, and who is in a position to consider matters such as shareability and security. Ownership does not imply any rights to use or disposal. During the period that the research project is under way it is likely that the owner will be a research officer or an individual in a supervisory rôle. At project end the ownership should be transferred to an appropriate individual, such as the project PI or the data manager responsible. In many cases it will be appropriate for a research officer to retain ownership.

Confidentiality Status

Confidentiality status indicates what classes of people and what automated information-gathering systems may have sight of the data record; it does not provide information about how such records are protected. It is likely that the confidentiality status will change during the life-cycle of the data record, in which case the status must be updated. Access is either free or limited. If access is free, then the term 'public domain' should be used. If the access is limited, then the entities who are permitted to see this data should be identified either by naming groups or individuals.

Record Title	File Name	Owner	Contact Details	Data Record Type	Confidentiality Status
Example:					
IdMRC Research Project Data Record Manifest	*erim6mar11021.mpl*	*Mansur Darlington*	*ensmd@bath.ac.uk*	*associative data record*	*public domain*

History of this PDRM

Figure 2: Project Data Record Manifest Template for IdMRC Projects

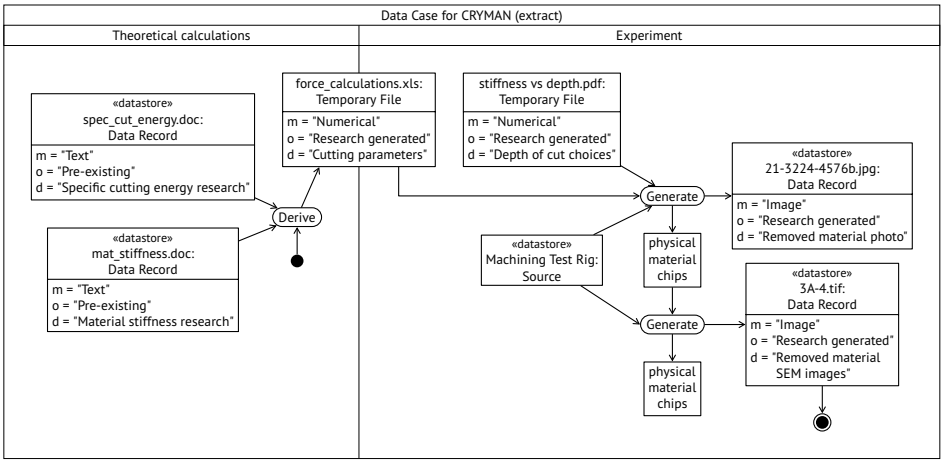


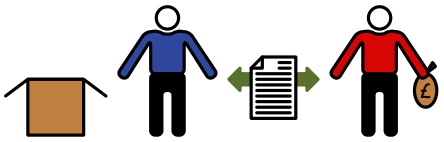
Figure 3: Example RAID diagram

The main component of this (let me zoom in ¶) is a table listing all the records associated with a project, showing the record title, file name and location, owner and contact details, record type, and confidentiality status. With this at least we have a chance of working out what's what. But filling out that table is laborious, and very easy to forget to do until the task becomes monumental. Plus, there's a lot a mere table can't do, such as indicate which files derived from which other files. So we came up with the idea of a RAID diagram. ¶

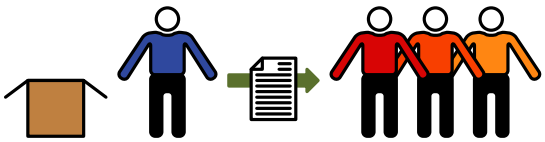
This (Figure 3) is an extract from a RAID diagram we did for an investigation into machining cryogenically frozen materials. You can see fairly clearly where the machining parameters came from and which runs these two images came from. It's better, but still a bit cumbersome to do by hand, so we put together a tool called RAIDmap to make it dead easy. ¶ It is based on the Open University's Compendium tool, with a few bells and whistles added, and while it's a bit rough around the edges it is available for anyone to download and use. ¶

I was also involved in a project called the Smart Research Framework. It is developing tools that go several steps further by building data management directly into researcher

- Contracts



- Pure licences



- Waivers

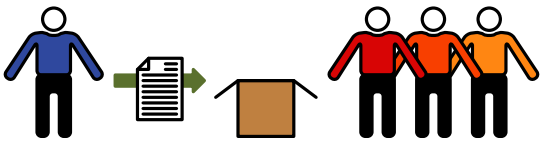


Figure 4: Types of licences

workflows. One of the tools is an electronic lab notebook system called LabTrove. It's based on blog technology, where each post can represent a sample, a technique, a methodological stage, or the data output from a particular run, and they all link together to make a highly efficient and easily navigable scientific record. Standalone instances of this technology have been installed in various places and it can also be used as a cloud application.

There are lots of other relevant tools I could mention here, such as Manchester's MADAM environment, or Oxford's DataFlow products, but I really must move on to my last two challenges, which are both about removing barriers to reuse. The first is one that is easy for researchers to sort out, but next to impossible for anyone else; and that is ¶ licensing.

4 Licensing data

Intellectual property law as it relates to data is hard, not least because different aspects of the data – keys, values, structure, data model, derived visualisations – may all be treated quite differently. And as research is a global concern, there are real international differences to contend with: in the Europe we place a lot of emphasis on sweat of the brow, in Australia they value originality, and in the US creativity. All of which makes the default legal position terribly hard to uncover. This is where licensing comes in. It's a way for researchers to make it absolutely clear what can be done with their data. ¶

None of what I'm about to say is legal advice, as I am not a lawyer, but my understanding is that there are three ways of licensing data (Figure 4). The first is by contract, where two named parties exchange rights and responsibilities; as a rule of thumb, you can spot these by the fact you need to sign them. The second is by true licence, which you can attach to a resource and it will automatically apply whenever someone uses it. The last is by waiver, where you formally give up rights over the resource. ¶

So, the questions researchers need to ask themselves are:

1. Do I need to make a choice? A particular licensing arrangement may be mandated by
    - Institutional policy
    - Data archive policy
2. If so, would a standard licence suffice? The CC0 waiver gives the most flexibility but doesn't work in Australia. If reserving some rights, currently I think the Open Data Commons licences are the most suitable, though that will change when the Creative Commons version 4 licences are released.
3. If not, how do I write my own licence? Well, the most important thing is to get help from your Research Office and legal department. Typically you'd only need a custom licence if you need to control access to the data or usage of them in some way.
4. Do I need more than one licence? You might, for example, want to provide a non-commercial licence and sell a commercial one.

Actually applying the licence is a matter of making it visible anywhere a potential reuser is likely to look: on a dataset landing page, in a README or LICENSE file supplied with the data or embedded in file metadata. ¶

5 Making data citable

Lastly, I want to talk about how to make data citable. There are two aspects to this: one is about metadata and the other is about infrastructure. ¶ In terms of metadata, the bare minimum that any potential reuser will need to know about a dataset in order to cite it is:

- Author
- Date made available
- Title
- Publisher/host
- Location (= identifier)

None of these is quite as straightforward as they seem. People don't usually write data, they collect them, so 'author' doesn't really tell you whose name should go there. I can see a time where you'll list the P.I.s in the citation, and all the other contributors listed alongside their roles in the metadata, but we're quite a way off establishing best practice on that.

Typically the date will be the year when the data repository first published the landing page, but for dynamic datasets the date of last change might be more appropriate.

As for the title, I'd recommend making this different from that of the primary research paper, if only by putting 'Data from:' at the start.

The publisher or host gives another way of finding the copy of the data if, heaven forbid, the link should break, and might also lend the dataset some quality assurance.

Lastly the location for digital objects should be a link from which it should be downloaded, so a dataset landing page rather than the data itself. Landing pages are great for all sorts of reasons I don't have time to go into. And ideally the link should be in the form of a resolver service plus an identifier such as a DOI. Not only does this make it easier to move the data around without breaking any links, but you also get an identifier which can be used for tracking impact, or as a key in various different automated systems.

By the way, there is a general but not universal consensus that DOIs should always point to the exact same version of a dataset, so readers always see the version that the author saw. ARKs, Handles and PURLs are a better fit for datasets that are changeable, inaccessible or draft.

There are infrastructure implications to making all that information available to reusers. The last one in particular implies putting the data in a repository or archive that will look after it and keep it accessible. You just wouldn't get that from simply bunging the data up on a departmental web page. Plus, if the researcher is to get academic credit for a dataset, it needs to be somewhere that will be indexed by, say, Thompson Reuters' Data Citation Index. ¶

6 More Information

If you want more information on any of the topics I've talked about today, ¶ here are some useful guides.

DCC How-to Guides: `http://www.dcc.ac.uk/resources/how-guides`

- How to Cite Datasets and Link to Publications
- How to Develop a Data Management and Sharing Plan
- How to License Research Data
- How to Develop Research Data Management Services (in preparation)

ANDS guides: `http://ands.org.au/guides/#datamanagement`

- Creating a Data Management Framework
- Data Management Planning
- Ethics, Consent and Data Sharing
- Storage

Or you can of course, get in touch with me directly or ask me something now.

Alex Ball. DCC/UKOLN, University of Bath. `http://www.ukoln.ac.uk/ukoln/staff/a.ball/`

The DCC is funded by JISC.

For more information, please visit `http://www.dcc.ac.uk/`